Personalized Speech Enhancement through Self-Supervised Data Augmentation and Purification

Aswin Sivaraman, Sunwoo Kim, Minje Kim

Indiana University, Department of Intelligent Systems Engineering, USA

{asivara, kimsunw, minje}@indiana.edu

Abstract

Training personalized speech enhancement models is innately a no-shot learning problem due to privacy constraints and limited access to noise-free speech from the target user. If there is an abundance of unlabeled noisy speech from the test-time user, one may train a personalized speech enhancement model using self-supervised learning. One straightforward approach to model personalization is to use the target speaker's noisy recordings as pseudo-sources. Then, a pseudo denoising model learns to remove injected training noises and recover the pseudo-sources. However, this approach is volatile as it depends on the quality of the pseudo-sources, which may be too noisy. To remedy this, we propose a data purification step that refines the self-supervised approach. We first train an SNR predictor model to estimate the frame-by-frame SNR of the pseudo-sources. Then, we convert the predictor's estimates into weights that adjust the pseudo-sources' frame-by-frame contribution towards training the personalized model. We empirically show that the proposed data purification step improves the usability of the speaker-specific noisy data in the context of personalized speech enhancement. Our approach may be seen as privacy-preserving as it does not rely on any clean speech recordings or speaker embeddings.

Index Terms: speech enhancement, self-supervised learning, privacy-preserving machine learning, model compression

1. Introduction

Speech enhancement is a well-studied research area within signal processing [1–3] which has experienced significant progress in the past decade due to the pervasiveness of machine learning models and deep neural networks (DNNs) [4–8]. The majority of automated noise suppression algorithms introduced over the years are geared towards general-purpose ("universal" or "speaker-agnostic") speech enhancement. In this context, denoising models are trained to separate speech from noise without prior knowledge of the speaker identity or the noises present. However, given the proliferation of voice-controlled devices (e.g., smart headphones and smart speakers), we anticipate the need for "personalized speech enhancement" models that can optimize a single speaker's enhancement with respect to their unique acoustic environment.

Comprehensive studies of DNN-based speech enhancement or speech separation systems have shown that a model's generalization power depends on its complexity and architecture. It was shown, for example, that a large fully-connected DNN with 2048 units and 5 layers can generalize well to unseen noise sources [9] but may not adapt to unseen test speakers. Instead, a long short-term memory cell (LSTM) network achieved the generalization goal in speaker- and noise-agnostic separation tasks [10]. However, it still requires a substantially large network architecture (1024×4). Other studies have also shown that a mismatch between the training and test input signals may result in highly varied performance unless the model has been exposed to an excessive amount of data [11]. Mismatching factors include the type and loudness of the noise and the characteristics of the speaker. In summary, for a speaker-agnostic generalist model to optimally address the peculiarities of a particular test-time user and their environment, one must both increase the diversity of the training speech and noise corpora and increase the model complexity.

These increases in training data and model complexity induce a trade-off, as with personal devices, efficient test-time inference is of prime importance given to the often limited resources. In this paper, we address this trade-off by developing a specialist model. We define a specialist as a smaller model which solves a subset of the original problem intended for the generalist. As such, we can afford a reduction in the overall number of parameters. Recent research has explored the benefits of specialization towards speech enhancement in recent years. The VoiceFilter model informs the speech enhancement model of the estimated speaker-identifying information [12]. In [13], speaker- or gender-specific models outperform a generalist model. These studies did not utilize personalization as a manner of reducing model complexity; however, one study extends the idea to a mixture of local expert architecture, where the test-time specificity is identified and then assigned to a few pre-defined specialists, achieving model compression [14].

Another challenge in personalized speech enhancement is that it is not always possible to acquire clean speech data from the test-time user. For example, speech enhancement models within modern-day smart devices might be trained through always-on ambient data collection. This trend is at odds with user concerns regarding privacy and security [15]. A recent DNN-based system required as little as five seconds of clean speech data from a test-time speaker in order to convincingly synthesize new utterances out of the previously unseen speaker's voice [16]. Breakthroughs such as these may make users reluctant to provide any clean speech recordings to their smart devices. Realistically then, training a personalized model should be viewed as a no-shot machine learning task [17, 18]. While eliminating reliance on clean speech recordings from the test-time user may not fully remedy all privacy concerns, we believe speech enhancement models which minimize personal data collection are always desirable from the user's perspective.

In this paper, we take a less intrusive route to achieve personalization by using only noisy data from the test-time speaker. This setup exceeds the scope of a fully supervised formulation for training a denoising model, which typically requires pairs of artificial mixtures and clean reference signals. Instead, a self-supervised learning approach may be better suited; this works by optimizing the model based on a pretext task which proxies the intended task [19]. This paradigm has seen extensive usage in computer vision research [20, 21], with recent studies

applying the concept towards speaker-agnostic speech enhancement [22]; our paper uses self-supervised learning directly towards speaker-specific personalized speech enhancement.

To this end, we improve the quality of the test user's noisy data by incorporating a data purification step, as conceivably, some audio frames of the noisy speaker-specific dataset may contain more clean speech than others. Rather than considering every frame as equal, the self-supervised formulation may benefit from additional prior knowledge which emphasizes specific frames based on the presence of clean speech. Our proposed method introduces a weighting scheme derived from a frame-by-frame estimate of the noisy speech's signal-to-noise ratio (SNR). An explicit SNR prediction step has been used before to boost the performance of DNNs for speech enhancement in a fully supervised setup [23, 24]. However, our work is the first to apply this step in the context of personalized speech enhancement. By weighting the frames based on their SNR, we inexpensively label the unlabeled noisy data. This data purification can guide the speaker-specific self-supervised learning objective towards better approximating a hypothetical speaker-specific fully-supervised learning objective.

Our paper's contributions may be summarized as follows: (1) we formulate the personalized speech enhancement context, whose training is done using noisy data of the intended test-time speaker rather than the clean voice; (2) we introduce one method of self-supervised learning for personalized speech enhancement, which treats the noisy speaker data as pseudo-sources; (3) we propose a data purification step which modifies the self-supervised learning loss function to weight the contributions of the noisy pseudo-source training data based on the frame-by-frame "cleanliness score", or SNR.

By avoiding explicit calculation of any speaker-identifying embedding vectors and without using any clean speech data, we assert that the proposed methods are first steps towards privacy-preserving personalized speech enhancement.

2. Methods

2.1. Fully-Supervised Speech Enhancement

Speech enhancement (SE) is commonly posed as a fully supervised learning problem, in which a model learns to map noisy mixture signals to clean speech signals by processing pairs of inputs and targets. The input mixtures $\boldsymbol{x}$ are made by artificially mixing clean speech utterances $\boldsymbol{s}$ with training noise signals $\boldsymbol{n}$; the amplitude of $\boldsymbol{n}$ may be scaled to simulate various SNRs. The utterances are sampled from a large training dataset containing many speakers, $\boldsymbol{s} \in \mathsf{S}$, and the noises from a similarly large dataset of diverse noises, $\boldsymbol{n} \in \mathsf{N}$. The denoising model g updates its parameters $\mathcal{W}_g$ with each iteration such that the distance $\mathcal{E}$ between the denoised estimate signal $\boldsymbol{y}$ and the target clean speech signal $\boldsymbol{s}$ is minimized. The learning procedure for the generalist model may be summarized as follows:

$$\text{Mixture: } \boldsymbol{x} = \boldsymbol{s} + \boldsymbol{n}; \quad \boldsymbol{s} \in \mathsf{S}; \quad \boldsymbol{n} \in \mathsf{N} \tag{1}$$

$$\text{SE Objective: } \underset{W_g}{\operatorname{argmin}} \quad \mathcal{E}(\boldsymbol{y} = g(\boldsymbol{x}; \mathcal{W}_g) \parallel \boldsymbol{s}) \tag{2}$$

There are many potential choices for the loss function $\mathcal{E}$—in this study, we utilize time-domain mean square error ($\mathcal{L}_{\text{MSE}}$), which is the per-sample squared distance between the estimate ($\boldsymbol{y}$) and target ($\boldsymbol{s}$) waveform pairs of length L,

$$\mathcal{L}_{\text{MSE}}(\boldsymbol{y}; \boldsymbol{s}) = \frac{1}{L} \sum_{i=0}^{} (\boldsymbol{s}_i - \boldsymbol{y}_i)^2 : \tag{3}$$

MSE has been shown to correlate well with improving the objective signal quality [25], but denoising performance is commonly reported in scale-invariant signal-to-distortion ratio (SI-SDR) [26].

A naïve approach to personalized speech enhancement would be to replicate this procedure using only speaker-specific data. But because we consider the personalized speech enhancement to be a no-shot learning problem, our study assumes that we do not have access to ground-truth clean speech utterances. Therefore, the conventional fully-supervised learning objective cannot be used directly in training a personalized speech enhancement model.

2.2. Self-Supervised (Pseudo) Speech Enhancement

We assume that the test-time speaker's easily collected noisy speech data can be described as a mixture of their clean utterances corrupted by a set of unknown additive noises, $\tilde{\mathsf{S}}^{(k)} = \mathsf{S}^{(k)} + \mathsf{M}$. We can simulate this "premixture" process in our experiments by sampling utterances from the test-time speaker, $\boldsymbol{s}^{(k)} \in \mathsf{S}^{(k)}$, and mixing them with a corpus of designated premixture noises, $\boldsymbol{m} \in \mathsf{M}$; the model is only allowed to access the premixed signals $\tilde{\boldsymbol{s}}$ and not its components $\boldsymbol{s}^{(k)}$ and $\boldsymbol{m}$. From here, we omit the speaker index, superscript k, for brevity.

Our proposed self-supervised learning strategy treats the premixtures as the new set of training targets. Therefore, the premixtures $\tilde{\boldsymbol{s}}$ are injected with further training noises, $\boldsymbol{n} \in \mathsf{N}$, to create a new set of input mixtures $\tilde{\boldsymbol{x}}$. The self-supervised model f_{PSE} updates its parameters $\mathcal{W}_{f_{\text{PSE}}}$ by mapping the doubly-corrupted input mixtures $\tilde{\boldsymbol{x}}$ to a pseudo-denoised estimate signal $\tilde{\boldsymbol{y}}$ and minimizing its distance to the originating premixture source $\tilde{\boldsymbol{s}}$—in other words, the self-supervised model learns to undo only the second noise injection.

The two discussed mapping functions are non-equivalent, i.e., $f_{\text{PSE}} \neq g$, not only because g is trained on data from many speakers while f_{PSE} is trained on data from a single speaker, but also because f_{PSE} is trained using non-clean source signals which makes it a pseudo speech enhancement (PSE) model. Suppose there exists a hypothetical optimal speaker-specific denoising function (f); our hypothesis is that f_{PSE} better approximates f as opposed to the fully-supervised speaker-agnostic generalist function g.

Because it does not directly solve the speech enhancement problem, while it still mimics the source-separating nature via data augmentation on unlabeled signals, we consider our proposed learning objective analogous to being a pretext task. The self-supervised training procedure summarized as follows:

$$\text{Premixture: } \tilde{\boldsymbol{s}} = \boldsymbol{s} + \boldsymbol{m}; \quad \boldsymbol{s} \in \mathsf{S}^{(k)}; \quad \boldsymbol{m} \in \mathsf{M} \tag{4}$$

$$\text{Mixture: } \tilde{\boldsymbol{x}} = \tilde{\boldsymbol{s}} + \boldsymbol{n}; \quad \boldsymbol{n} \in \mathsf{N} \tag{5}$$

$$\text{PSE Objective: } \underset{W_{f_{\text{PSE}}}}{\operatorname{argmin}} \quad \mathcal{E}(\tilde{\boldsymbol{y}} = f_{\text{PSE}}(\tilde{\boldsymbol{x}}; \mathcal{W}_{f_{\text{PSE}}} \parallel \tilde{\boldsymbol{s}}) \tag{6}$$

Figure 1 illustrates the impact of the two stages of noise which are applied to the clean speech waveform. In short, our specialist models trained using pseudo speech enhancement are optimized in the same manner as the generalist models, by minimizing the per-sample distance between pairs of estimates and targets, however the target in this case is a pseudo-source, i.e. $\mathcal{E}(\tilde{\boldsymbol{y}} \,||\, \tilde{\boldsymbol{s}}) = \mathcal{L}_{\text{MSE}}(\tilde{\boldsymbol{y}}; \tilde{\boldsymbol{s}})$. This approach bears similarity to the recently proposed mixture invariant training (MixIT) procedure [27].

Figure 1: Illustration of the audio data transformations through pseudo speech enhancement (PSE). Premixtures represent real-world noisy audio recordings from the test-time speaker, and are the training target of pseudo-denoising model f_{PSE}.

2.3. Data-Purified Pseudo Speech Enhancement

The success of deriving meaningful speaker-specific features from pseudo speech enhancement depends on the quality of the premixture—more specifically, the sparsity of m in time, as well as the instantaneous SNR between s and m, are both factors as to whether s is too degraded to be usable. If m is sufficiently sparse, portions of the premixture may contain near-clean speech. Our goal is to inform the enhancement model of where these near-clean frames may be. We propose introducing "data purification" (DP) to pseudo speech enhancement training; the data-purified model f_{DP} estimates a weighting vector p which diminishes the contribution of contaminated frames towards the loss function L. By masking out the noisy premixture frames, the personalized model will hypothetically learn only using snippets of clean speech from the test-time user. This will differ from the original self-supervised model, i.e. $f_{DP} \neq f_{PSE}$, but our hypothesis is that it may better approximate the ideal denoising function, i.e. $f_{DP} \approx f$.

Our method for generating p is to train a separate model h which can estimate the segmental SNR of the premixtures, calculated over a set of windowed overlapping frames. The SNR predictor is a regressive model, trained over the diverse set of training speakers and noises (i.e. G, and N), which outputs a vector of instantaneous SNRs $\hat{\gamma}$; it has no knowledge of the test-time speaker or the test-time noise environment. Given an estimate signal $\hat{v}$ and a target signal v, both of length L, their residual is $r = v - \hat{v}$, and the frame-by-frame/segmental SNR (SegSNR) can be defined as:

$$\text{SegSNR}_j(v, \hat{v}) = 10 \log_{10} \left[\frac{\sum_{i=Hj}^{Hj+N-1} (w_i v_i)^2}{\sum_{i=Hj}^{Hj+N-1} (w_i r_i)^2} \right]; \quad (7)$$

where N is the frame size, H is the hop size, j is a zero-based frame index (i.e. $0 \leq j \leq d\frac{L}{H}e - 1$), and vector w comes from the Hann window function of length N. Note that the SNR predictor inputs are of length L and outputs are of length $d\frac{L}{H}e$. Its training objective may then be summarized as:

$$\text{Mixture: } x = s + n; \quad s \in G; \quad n \in N \quad (8)$$

$$\text{Target: } \gamma = \text{SegSNR}(s, x) \quad (9)$$

$$\text{SNR-Predictor Objective: } \underset{W_h}{\operatorname{argmin}} E(\hat{\gamma} = h(x, W_h) - \gamma); \quad (10)$$

When training the pseudo-denoising model, the fully-trained SNR predictor first analyzes input premixtures to estimate the instantaneous SNRs, $\hat{\gamma} = h(s)$; we apply the logistic function σ to the $\hat{\gamma}$ logits to obtain frame-by-frame weights:

$$p = \sigma(\hat{\gamma}) = \frac{1}{1 + e^{-\hat{\gamma}}} \quad (11)$$

Figure 2: Illustration of the fully-trained SNR predictor inputs and outputs. The first subplot features an example premixture / pseudo-source s. In the second subplot, the SNR predictor network h estimates the instantaneous SNR of the premixture. The third subplot shows $\hat{\gamma}$ converted to weights using the logistic function, i.e. $p = \sigma(h(s))$.

The training procedure for the data purified pseudo speech enhancement (PSE + DP) model mirrors Eq. (4)–(6) except that we modify the loss function to now incorporate the frame-by-frame weighting vector through a custom segmental MSE function, i.e. $E(y - k\, s) = L_{\text{SegMSE}}(y, s, p)$, where

$$L_{\text{SegMSE}} = \frac{1}{J} \sum_{j=0}^{J-1} p_j \left[\frac{1}{N} \sum_{i=Hj}^{Hj+N-1} (w_i s_i - w_i y_i)^2 \right]. \quad (12)$$

Here J is the number of frames $d\frac{L}{H}e$. The mean-squared difference is taken between the windowed segments, which are then weighted by p then averaged across all frames.

3. Experiment Setup

3.1. Configurations

Our experiment considers a baseline and four proposed training procedures in potentially developing personalized speech enhancement models. We group the proposed methods based on whether we pretrain the models using random initialization or speaker-agnostic fully-supervised pretraining. Regardless of initialization, all the proposed configurations use pseudo speech enhancement as the self-supervised learning approach to personalization. We additionally examine the impact of the proposed data purification scheme.

- SE: Trained to minimize Eq. (2). This is our generalist baseline, the speaker-agnostic speech enhancement system.
- PSE: The proposed plain pseudo speech enhancement method. This self-supervised learning method relies solely on noisy speaker-specific data to minimize Eq. (6).
- PSE+DP: A self-supervised setup using Eq. (6). However, the model uses the weighted segmental MSE L_{SegMSE} instead Eq. (12) to purify the noisy speaker-specific dataset.
- SE $\rightarrow$ PSE: Instead of random initialization, a model is first trained to minimize Eq. (2), then fine-tuned to minimize Eq. (6).
- SE $\rightarrow$ PSE+DP: Same as above, but with data purification.

3.2. Data Preparation

We opt for an online data augmentation procedure which combines three different public audio datasets (Librispeech

[28], MUSAN [29], and FSD50K [30]) to test our methods' robustness cross-dataset. Our speaker-specific datasets $S^{(k)}$ are stochastically sampled out of all utterances from one test speaker. This is done for twenty folded-out test speakers. The speaker-agnostic dataset S is stochastically sampled out of all utterances across the remaining 211 speakers within the Librispeech train-clean-100 split. The audio clips from the FSD50K dev split serve as the premixture noises M. The training noises N come from the MUSAN free-sound split. Model performance is evaluated on a set of random mixtures between unseen utterances within the test-time speaker dataset combined with unseen noises from the MUSAN sound-bible split.

During training, for any given mixture, the sampled signals (s, m, n) are 1 sec in length with a sample rate of 16 kHz. Each premixture noise m is scaled uniformly at random such that the premixture SNR falls between 0 to 15 dB, whereas the training noises n are scaled uniformly at random such that the mixture SNR falls between 5 to 5 dB. Our decision of premixture SNR range is based on real-world scenarios, e.g., a smart speaker collecting noisy speech data in the living room.

3.3. Models

Our experiment focuses on the correlation between the number of effective model parameters and the improvement in speech enhancement quality (SI-SDR) obtained through the SNR-informed data purification method. Because our real-world use case relates to compute-constrained smart devices likely to perform low-latency speech enhancement on-device, we evaluate small neural network architectures which do not compete with popular speech separation models, e.g., Conv-TasNet [31], or Dual-Path RNN [32]. Our hypothesis is that low-capacity models will benefit the most from personalization, as they do not generalize well.

We compare three two-layer GRU-based models (varying hidden units: 64, 128, and 256) so as to compare the relationship between the five training configurations and increased model complexity. These models perform denoising via time-frequency masking [33]—audio waveforms are converted to the time-frequency domain using the short-time Fourier Transform (STFT) with a frame size of $N = 1024$ and a hop size of $H = 256$; as the waveforms are 16000 samples in length (1 sec), this results in $J = 63$ STFT frames. The denoised signal is obtained by applying the estimated time-frequency mask onto the input STFT and performing the inverse transform. The loss is computed in the time domain.

The SNR predictor model uses a GRU with 1024 hidden units and 3 hidden layers; it processes STFTs with the same N and H, performing a frame-by-frame regression. These choices of N, H, and J apply towards Eq. (7) and (12). We defer the investigation of smaller SNR predictor networks intended for on-device training to future work.

4. Results

Fig. 3 summarizes the results of our experiment. We can observe the model compression benefits of personalized speech enhancement by comparing the specialist models with fewer parameters with the generalist models with more parameters. E.g., a 64 hidden unit network, trained using GE! PSE+DP, slightly exceeds the average denoising performance of a generalist model SE with 128 hidden units—this is an effective 59 % reduction in model size (from 412 k to 169 k parameters).

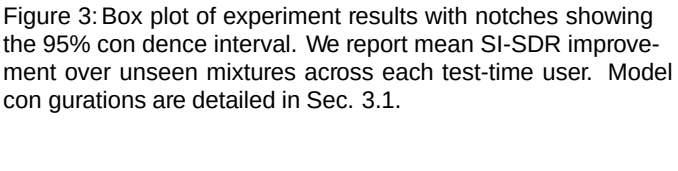

Figure 3: Box plot of experiment results with notches showing the 95% confidence interval. We report mean SI-SDR improvement over unseen mixtures across each test-time user. Model configurations are detailed in Sec. 3.1.

The personalized models trained with the five pretext task, PSE, underperform compared to the baseline non-personalized SE models with an equivalent number of parameters. This indicates that the configuration of our experiment—a realistic premixture SNR range of 0 to 15 dB—challenges the self-supervised model with obtaining speech denoising features out of noisy speech. As hypothesized, data purification overcomes this difficulty by ignoring the too-noisy frames. We see across all model sizes that the data purified self-supervised models PSE+DP consistently outperform equivalently sized baselines.

Our experiments show that the generalist-initialized specialists outperform the randomly-initialized specialists only marginally, at most by 0:22 dB. This suggests that the large multi-speaker corpus G and the model trained from it are limited in their ability to address all the peculiarities of our 20 test-time speakers. In other words, the denoising models f_{PSE} and f_{DP} do not approximate one another.

Lastly, we see that the smallest personalized GRU model benefits the most from self-supervised learning with data purification, e.g., SE! PSE+DP outperforms SE by 0:91 dB, while the largest GRU model gains the least, e.g., SE! PSE+DP outperforms SE by 0:34 dB.

5. Conclusion

This work introduced personalized speech enhancement as a no-shot learning problem which motivated a self-supervised learning solution. Our method treated noisy data as pseudo-sources. We personalized speech denoising models for twenty different speakers using three neural network architectures with varied model complexity. We compared a speaker-agnostic fully-supervised model against two proposed self-supervised models: one without and another with the proposed data purification that suppresses the contribution of low-SNR frames to the learning objective. Our study showed that the smallest models improved the most from personalization, for which the data purified self-supervised learning scheme yields the best denoising performance. Audio examples and source code are available at: https://saige.sice.indiana.edu/research-projects/pse-ssl-dp

6. Acknowledgement

This material is based upon work supported by the National Science Foundation under Grant No. 2046963.